同济大学交通运输工程学院
COLLEGE OF TRANSPORTATION ENGINEERING
TONGJI UNIVERSITY

交通数据分析

第九讲 分类：决策树、SVM

沈煜 博士 副教授
嘉定校区交通运输工程学院311室
yshen@tongji.edu.cn

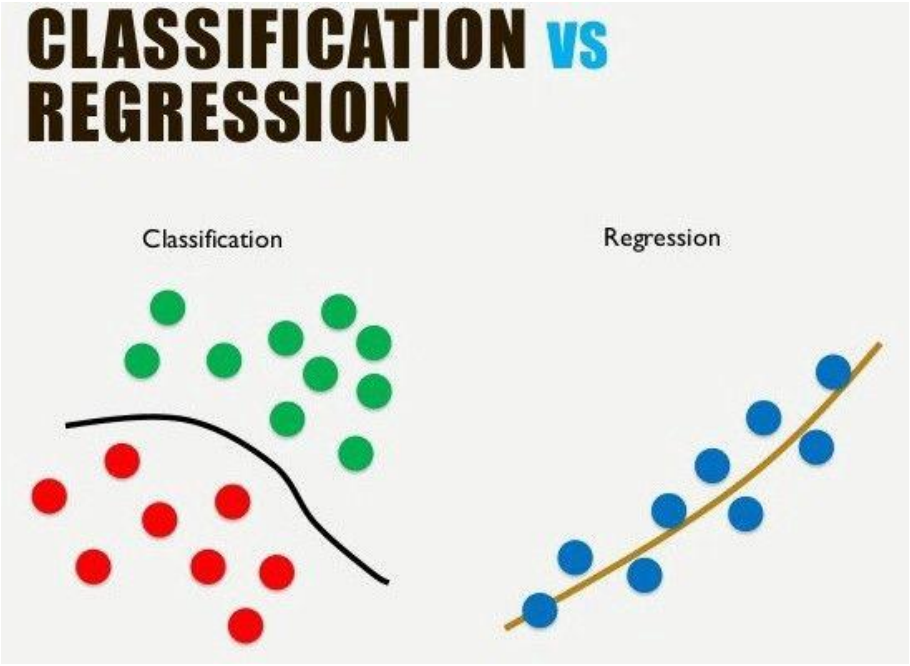
2022年04月22日

计划进度

周	日期	主讲	内容	模块
1	2022.02.25	沈煜	概述	
2	2022.03.04	沈煜	在线数据采集方法	爬虫
3	2022.03.11	沈煜	线性回归模型	回归分析
5	2022.03.18	沈煜	广义线性回归	
4	2022.03.25	沈煜	广义线性回归（**作业1**）	
6	2022.04.01	沈煜	空间数据描述性分析	
7	2022.04.08	沈煜	空间自回归方法（**作业2**）	
8	2022.04.15	沈煜	关联：Apriori	机器学习
9	2022.04.22	沈煜	决策树、支持向量机（**作业3**）	
10	2022.04.29	沈煜	浅层神经网络	
11	2022.05.06	沈煜	卷积神经网络（**期末大作业**）	
12	2022.05.13	沈煜	经典网络结构	
13	2022.05.20	沈煜	聚类：K-Means、DBSCAN	
14	2022.05.27	沈煜	贝叶斯方法、卡尔曼滤波	
15	2022.06.03	-	端午节放假	
16	2022.06.10	沈煜	期末汇报（1）	
17	2022.06.17	沈煜	期末汇报（2）	

内容纲要

- ➢基本概念
- ➢决策树归纳
- ➢支持向量机



基本概念

分类

分类 & 聚类

- **分类（有监督学习）**
  - 监督：**训练数据（观察，测量等）都带有标签**，指示观察的类别
  - 根据训练集分类新数据

- **聚类（无监督学习）**
  - 无监督：**训练集的类别（标签）未知**
  - 给定一个观察、测量等的数据集合，发现数据中存在的类或簇

分类 & 数值预测

- **分类与数值预测的区别**
  - **分类**：预测分类的**类标签**(离散 or 名义)，基于训练数据和类标签构造一个模型，并**分类新数据**
  - **数值预测**：构建连续数值**函数/模型**, **预测未知值/缺失值**
- **分类的典型应用**
  - 信用卡/贷款审批
  - 医疗诊断：肿瘤是癌或良性?
  - 欺诈检测：交易欺诈?
  - 网页分类：这是哪一类?
  - 交通：二元离散变量的分类预测（例如，通过/停止、交通事件有/无等）

分类：两步过程

- **模型构建：描述一组预先定义的类（分类训练）**
  - 假定每个元组/样本 属于一个类, 由类标签属性设定
  - 用于构建模型的元组集合称为训练集（training set）
  - 模型可以表示为分类规则,决策树, 数学公式
- **模型使用：分类将来/未知的对象（分类预测）**
  - **估计模型的准确率**
    - **测试集**：独立于训练集的样本 (避免过分拟合overfitting)
    - 比较测试样本的已知标签/由模型预测（得到）标签
    - **准确率**：测试样本集中模型正确预测/分类的样本的比率
- 如果准确率合适时，使用模型来分类未知样本（标签）

第一步：模型构建

训练数据

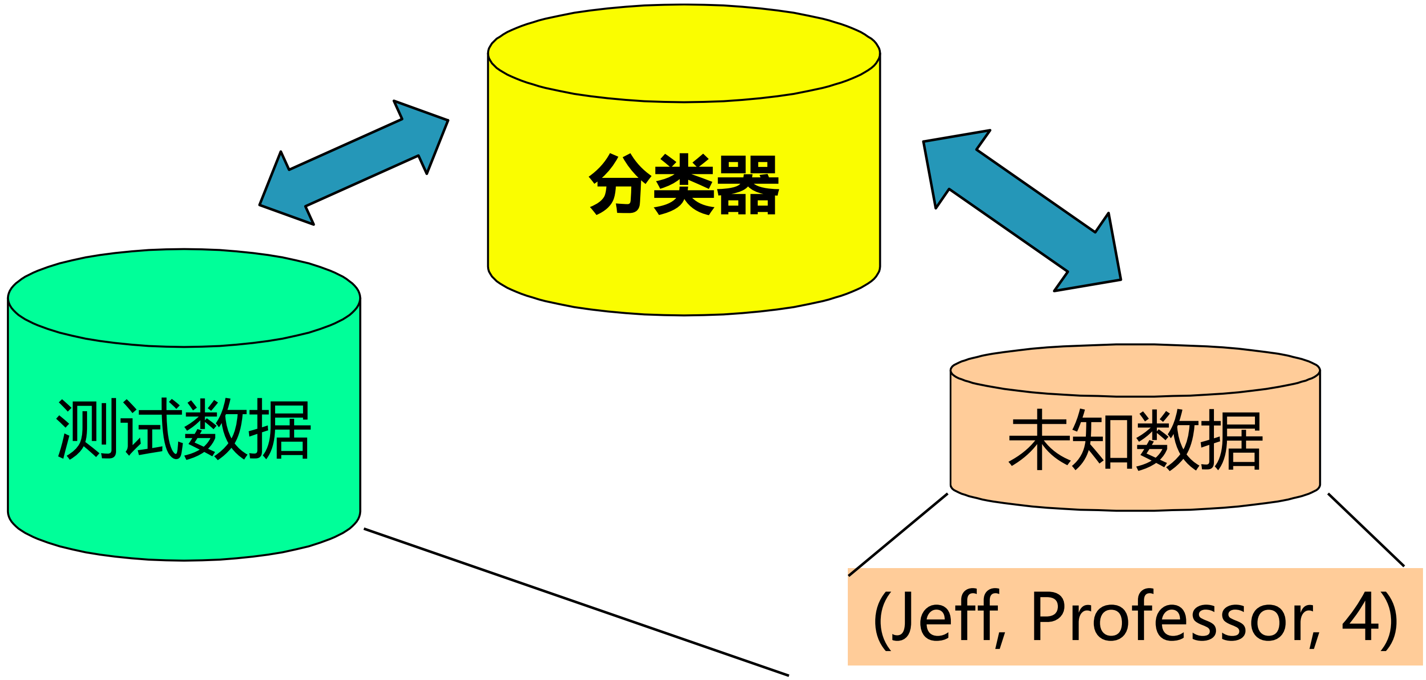
分类算法

分类器
(模型)

NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

IF rank = ‘professor’
OR years > 6
THEN tenured = ‘yes’

第二步：利用模型做出预测



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

Tenured?

Yes

如何评估分类方法的好坏?

- **准确性**
  - 分类准确性：是否对分类做出正确预测
  - 预测准确性：是否对变量的值做出正确预测
- **计算速度**
  - 训练时间：建立（训练）模型的时间
  - 分类/预测时间：使用模型进行运算的时间
- **鲁棒性**：处理噪音与缺失数据的能力
- **伸缩性**：对磁盘数据库进行处理的效率
- **解释性**：通过模型理解、洞察问题
- **其他度量方法**
  - 规则的优度（goodness）：决策树的大小、分类规则的紧致性

决策树

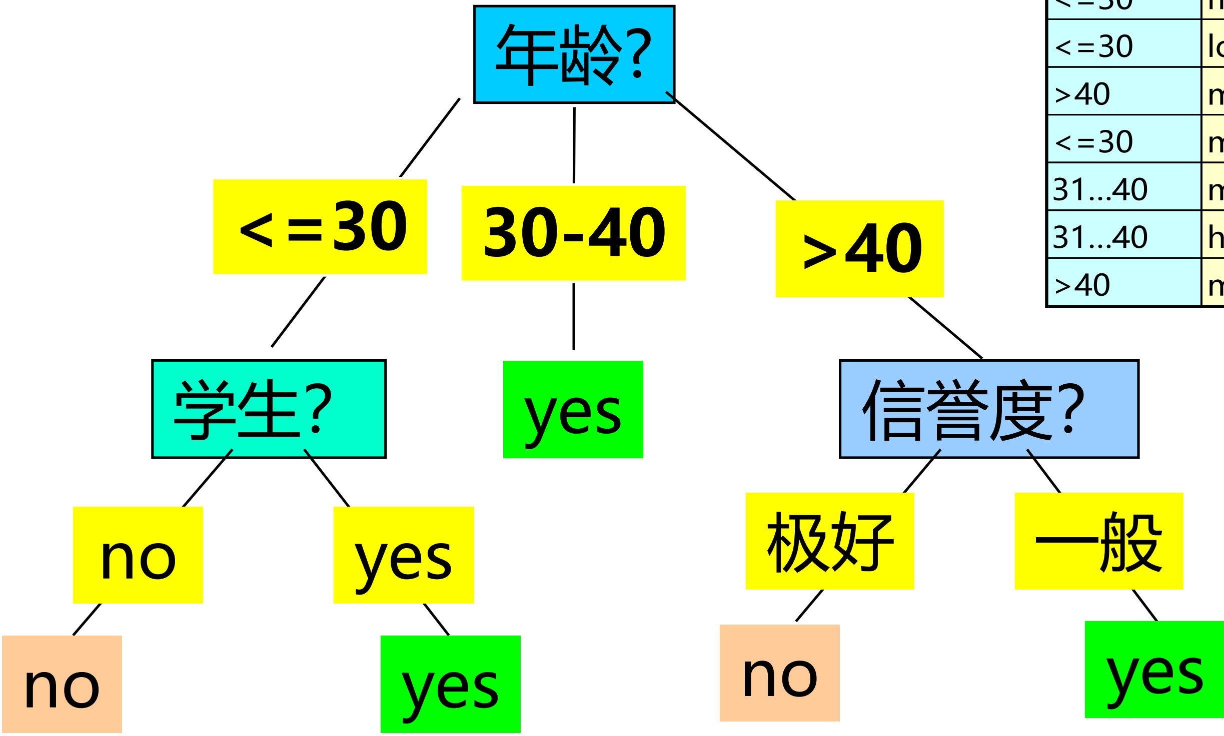
Decision Tree

决策树归纳: 例子

➢训练集: 购买计算机

年龄	收入	是否学生	信誉	购买计算机
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

➢结果:



决策树归纳的算法

贪心算法是指在对问题求解时，总是做出在当前看来是最好的选择。也就是说，不从整体最优上加以考虑，所做出的是在某种意义上的局部最优解。

- **基本算法（贪心算法）**
  - 树构建：自顶向下递归的分割方式
  - 开始阶段，所有的训练样本位于根节点
    - 属性是分类属性（若是连续值，事先离散化）
  - 基于选择的属性，样本被递归地分割
  - 基于启发式/统计测量来选择**分裂规则**（例如：信息增益、基尼指数等）
- **终止划分的条件**
  - 一个给定节点的所有样本属于一个类别
  - 没有属性剩下，用于进一步划分：运用多数投票来标记此节点
  - 没有样本剩下

伪代码

- **输出：一颗决策树**
- **方法：**
  - 创建一个节点N
  - **if** D中的样本都是同一类C
    - 返回N作为叶节点，以类C标记
  - **if** attribute_list为空 **then**
    - 返回N作为叶节点，标记为D中的多数类；//多数表决
  - 使用attribute_selection_method(D, attribute_list)，找出“最好”的splitting_criterion（分割标准）；
  - 使用splitting_criterion标记节点N；
  - **if** splitting_attribute是离散值的 **and** 允许多路划分 **then** //不限于二叉树
    - attribute_list ← attribute_list - splitting_attribute //删除划分属性
  - **for** splitting_criterion的 **each** 输出j //划分样本并对每个划分产生子树
    - 设D_j是D中满足输出j的数据样本的集合；//一个划分
    - **if** D_j为空 **then**
      - 加一个树叶到节点N，标记为D中的多数类；
    - **else** 加一个有Generate_decision_tree(D_j, attribute_list)返回的节点到节点N；
  - **end for**
  - 返回N

属性选择度量

- **属性选择度量**
  - 分裂规则决定给定节点上的样本如何分裂
  - 具有最好度量得分的属性选定位分裂属性
- **三种度量**
  - 信息增益、增益率、基尼（GINI）指标
- **数学符号**
  - D为样本的训练集，样本属于m个不同的类$C_i(i=1,\ldots,m)$
  - $C_{i,}$ D是D中的C_i类的样本集合
  - $|C_{i,}$ D$|$和$|$D$|$分别表示各自的样本个数

熵（Entropy）

➢熵（信息论）

- ➢用于度量不确定性的随机变量
- ➢计算方法
  - ➢对于m个完全不同的离散随机变量Y

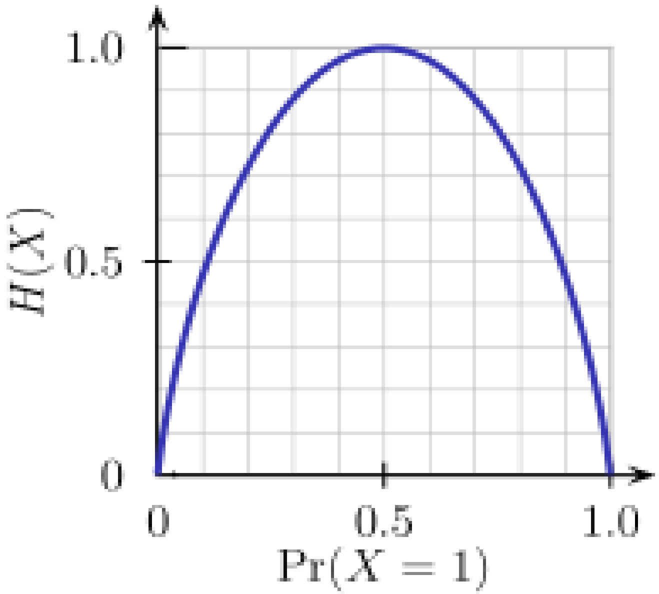
$H(Y) = -\sum_{i=1}^{m} p_i \log(p_i)$，

其中，$p_i = P(Y = y_i)$

- ➢解读
  - ➢高熵值：高不确定性
  - ➢低熵值：低不确定性

二元信息熵的四个特性

①非负性：H≥0
②对称性：对称于P=0.5
③确定性：P=0或1时，H=0
④极值性：P=0.5时，H最大，1



m=2

➢条件熵

- ➢$H(Y|X) = \sum_x p(x) H(Y|X = x)$

属性选择度量：信息增益(ID3/C4.5算法)

- **选择具有最高信息增益的属性**
- **令 p_i 为D中的任一样本属于类C_i概率，估计为 $|C_{i,} D|/|D|$**
- **分类D中样本需要的期望信息(熵，Entropy)：**
  - $\mathrm{Info}(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$
- **利用A分裂D为v个部分后，分类D需要的信息为**
  - $\mathrm{Info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \mathrm{Info}(D_j)$
- **以属性A分枝得到的信息增益**
  - $\mathrm{Gain}(A) = \mathrm{Info}(D) - \mathrm{Info}_A(D)$

属性选择：信息增益

- Class P：买电脑 = "yes"
- Class N：买电脑 = "no"
- $\text{Info}(D) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.940$
- $\text{Info}_{age}(D) = \frac{5}{14} \times \left(-\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right) + \frac{4}{14} \times \left(-\frac{4}{4}\log_2\left(\frac{4}{4}\right) - \frac{0}{4}\log_2\left(\frac{0}{4}\right)\right) + \frac{5}{14} \times \left(-\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right) = 0.694$
- $\text{Gain}(\text{age}) = \text{Info}(D) - \text{Info}_{age}(D) = 0.246$
- $\text{Gain}(\text{income}) = 0.029$
- $\text{Gain}(\text{student}) = 0.151$
- $\text{Gain}(\text{credit rating}) = 0.048$

年龄	收入	学生	信誉度	买电脑
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

年龄

- ☐ <30: Yes=2, No=3
- ☐ 31-40: Yes=4, No=0
- ☐ >40: Yes=3, No=2

Yes=9
No=5

计算信息增益-连续值属性

- **令 A 为连续属性**
  - 必须为A确定一个最佳分裂点（best split point）
- **上升序排序 A**
  - 一般，每对相邻值的中点是一个可能的分裂点
  - a_i和a_{i+1}之间的中点是$(a_i + a_{i+1})/2$
- **分裂点逐个计算，具有最小期望信息需求的点选为A的分裂点**
- **分裂**
  - D_1是满足 A ≤ split-point的样本集合
  - D_2是满足 A > split-point的样本集合

增益率 (C4.5算法)

- **信息增益倾向于有大量不同取值的属性（划分更细，更纯）**
  - 极端：每个划分子集只有一个样本，即一个类，此时$\mathrm{Info}(D) = 0$
- **C4.5（ID3 改进版） 使用增益率来克服这一问题（规范化信息增益）**
  - $\boxed{\text{Split Info}_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right)}$
  - $\text{Gain Ratio}(A) \ = \text{Gain}(A)/\text{Split Info}(A)$
  - 例如：$\text{Split Info}_{income}(D) = -\frac{4}{14}\log_2\left(\frac{4}{14}\right) - \frac{6}{14}\log_2\left(\frac{6}{14}\right) - \frac{4}{14}\log_2\left(\frac{4}{14}\right) = 1.557$
  - $\text{Gain Ratio}(income) \ = 0.029/1.557 = 0.019$
- **具有最大增益率的属性选为分裂属性**

Gini Index指标（CART算法）

- **Gini指标：如果数据D包含n类别的样本，那么，Gini(D) 定义为：**

$$gini(D) = 1 - \sum_{j=1}^{n} p_j^2$$

  - 其中，p_j为类别j在D中的频率

- **如果数据集D基于属性A分裂为子集D_1和D_2，那么，Gini指标定义为：**

$$gini_A(D) = \frac{|D_1|}{|D|} gini(D_1) + \frac{|D_2|}{|D|} gini(D_2\Big)$$

- **不纯度减少：** $\Delta gini(A) = gini(D) - gini_A(D)$

- **具有最小Gini split(D)的属性（或不纯度减少最大的）用于分裂节点：** 需要枚举所有可能的分裂情况

计算Gini Index指标

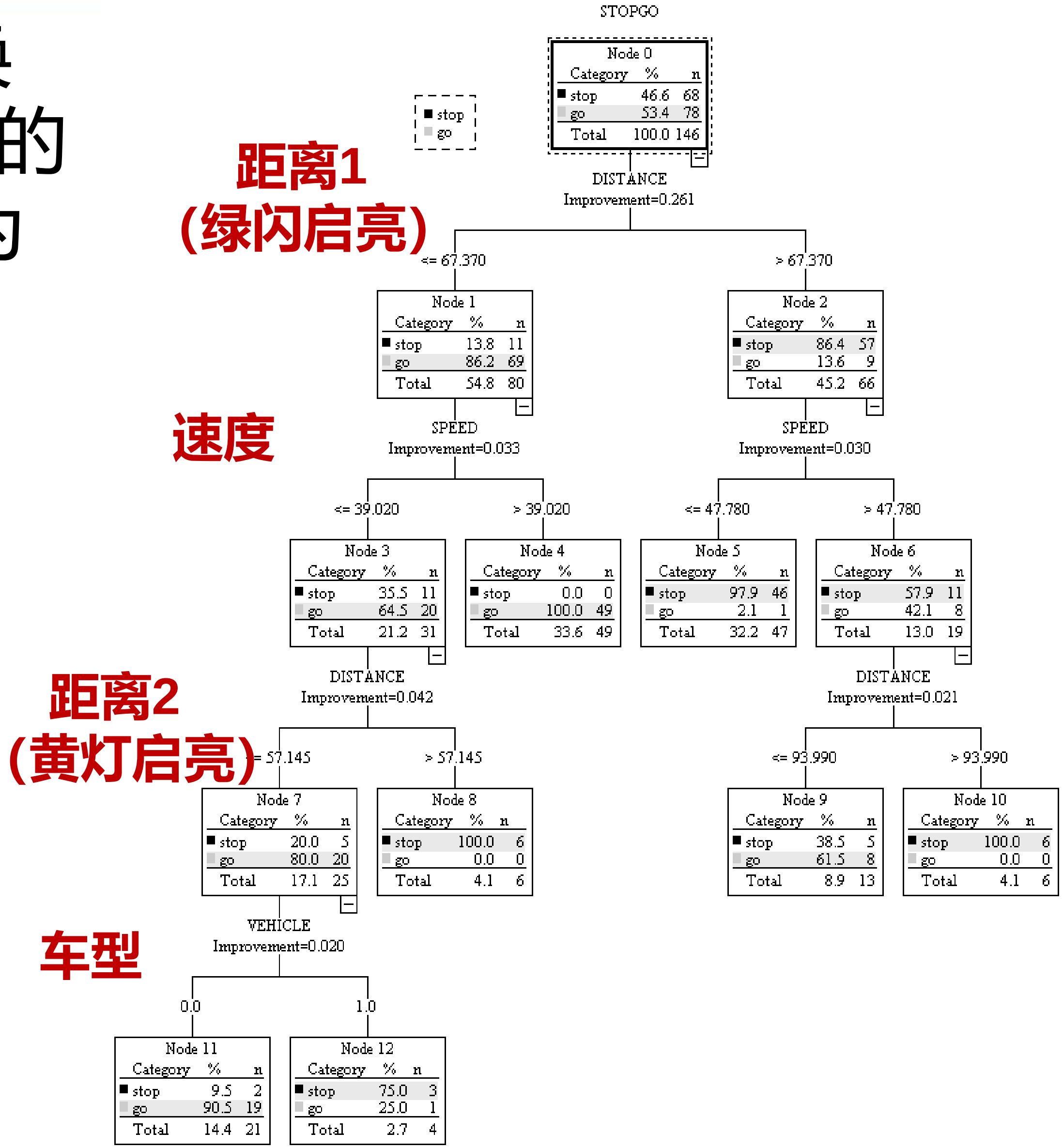
- D：有9个样本买电脑 = “yes”；有5个样本买电脑 = “no”
- 设属性income分裂D为包含10个样本的D_1：{low，medium}以及4个样本的D2 {high}
  - $Gini_{income\in\{low,medium\}}(D) = \frac{10}{14}Gini(D_1) + \frac{4}{14}Gini(D_2) = \frac{10}{14}\left(1 - \left(\frac{7}{10}\right)^2 - \left(\frac{3}{10}\right)^2\right) + \frac{4}{14}\left(1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right) = 0.443 = Gini_{income\in\{high\}}(D)$
  - 类似地，Gini{low, high} = 0.458; Gini{medium, high} =0.450
  - 由于具有最小的Gini index，因此{low, medium}、{high}分裂
- 假设所有属性都是连续值，需要其他技术（如聚类）来获得可能的分裂点

三种常用属性选择度量的比较

- **信息增益（Information gain）**
  - ➢偏向于多值属性
- **增益率（Gain ratio）**
  - ➢倾向于不平衡的分裂，其中一个子集比其他小得多
- **基尼指标（Gini index）**
  - ➢偏向于多值属性
  - ➢当类数目较大时，计算困难
  - ➢倾向于导致大小相等的分区和纯度

决策树在交通中的应用

➢驾驶员在相位切换（绿灯->红灯）时的停止-通过选择行为



支持向量机

线性最优分类超平面

基本原理

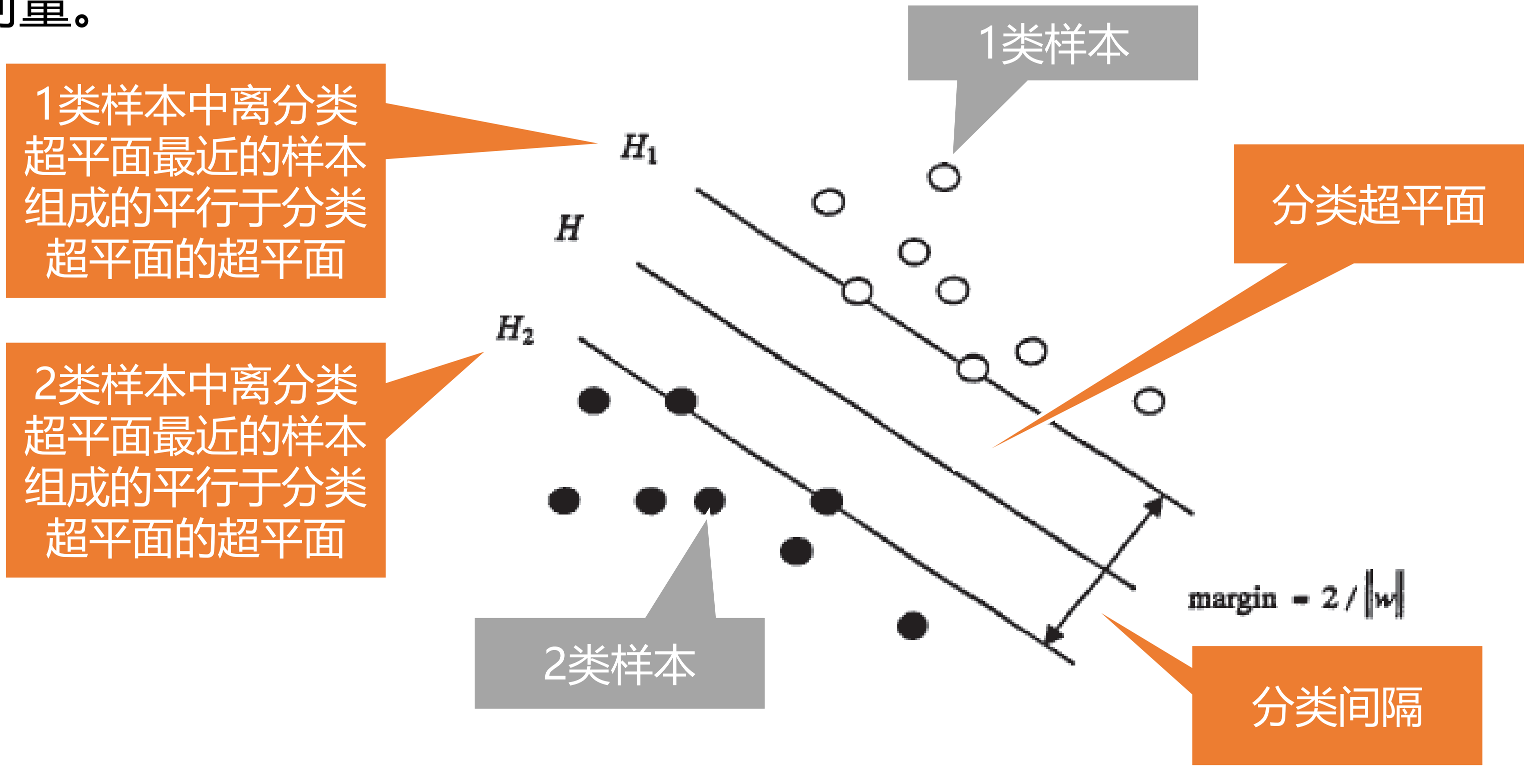
- **理论基础**
- 结构风险最小化Structural Risk Minimization (SRM)
- **基本思想**
- 通过某种特定的非线性映射把样本空间映射到一个高维乃至于无穷维的特征空间（Hilbert空间）
- 并在特征并在特征空间中寻求最优划分或回归线性超平面
- 把此平面作为分类决策面，从而解决样本空间中的高度非线性分类和回归等问题

主要特点

- 一种强有力的学习算法，在解决模式识别、回归估计等难题上具有良好效果
- 适用于小样本的机器学习问题，具有很强的数学理论基础
- 是一种可解析的机器学习方法
- 在深度学习出现之前，是最流行的机器学习算法之一

线性最优分类超平面

- 最优分类面不仅要求将两类正确分开，而且要使分类间隔最大。
- 将两类正确分开是为了保证训练错误率为0，也就是**经验风险最小**；使分类间隔最大实际上就是使推广性的置信范围最大，从而使**真实风险最小**。
- 将一维的分类**点**、二维的分类**线**、三维的分类**面**，推广到高维空间就是最优分类**超平面**，支持向量就是支撑这个超平面的不同维度上的数据点构成的向量。



线性最优分类超平面

- 假设两类**线性可分**的训练数据样本$\{(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)\}$，$x_i \in R^d$（d代表样本$x_i$的长度），$y_i \in \{-1,+1\}$，$i=1,2,\ldots,N$。
- 其**线性分类函数**的一般表达式是$f(x)=\omega*x+b$，该函数对应的**分类面H**为：

$$\omega * x + b = 0$$

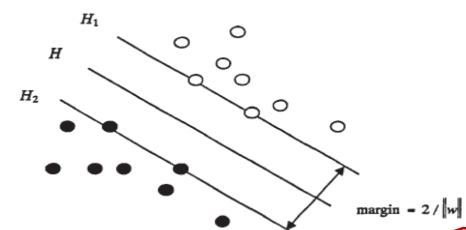
$$\begin{cases} \omega \cdot x_i + b \geq 1, y_i = 1 \\ \omega \cdot x_i + b \leq -1, y_i = -1 \end{cases}$$

第一类

第二类

(注意：二维空间是1个方程，多维空间则为1个方程组)

- 将**分类函数归一化**
- 使得对线性可分的样本集(x_i,y_i) $(i=1,2,\ldots,N)$
- 满足 $y_i[(\omega * x)+b]-1 \geq 0, \quad i=1,\ldots,N$



$\omega x + b = 0$

- 根据线性最优分类超平面的定义，**分类间隔**可表示为：

$$\rho = \min_{\{x_i,y_i=1\}} \frac{|\omega \cdot x_i + b|}{\|\omega\|} + \min_{\{x_i,y_i=-1\}} \frac{|\omega \cdot x_j + b|}{\|\omega\|} = \boxed{\frac{2}{\|\omega\|}}$$

线性最优分类超平面

- 要使分类间隔2/‖ω‖最大，等于使 $\|\omega\|^2/2$或者‖ω‖/2最小
- 这样线性SVM的**最优化分类面问题**可以表示成如下的**约束优化问题**：

目标函数 $$\min \Phi(\omega)=\frac{1}{2}\|\omega\|^2$$

约束条件 $$y_i[(\omega * x)+b]-1\geq 0,\quad i=1,\ldots,N$$

$$L(\omega,b,a)=\frac{1}{2}\|\omega\|^2-\sum_{i=1}^{N}a_i[y_i(\omega * x_i+b)-1]$$

(a_i≥0，拉格朗日乘子)

对ω，b，a 分别求导

$$\begin{cases}\dfrac{\partial L}{\partial \omega}=0 \Rightarrow \quad \omega=\sum_{i=1}^{N}a_i y_i x_i \\ \dfrac{\partial L}{\partial b}=0 \Rightarrow \quad \sum_{i=1}^{N}a_i y_i=0 \\ \dfrac{\partial L}{\partial a}=0 \Rightarrow \quad [y_i(\omega * x_i+b)-1]=0\end{cases}$$

解出最优的 ω, a, b，并代入原方程

$$f(x)=\operatorname{sgn}\left\{\sum_{sv}a_i^* y_i(x_i,x)+b^*\right\}$$

最优分类函数

广义最优分类面

- 非线性最优分类面
- 有少数样本使得原来线性可分的问题变成**线性不可分**问题，从而影响了分类器的性能，可以在约束条件中加入一个松弛因子ξ，即：

约束条件

$$y_i[(\omega * x)+b]-1\geq 0, \quad i=1,\ldots,N$$

$$\Rightarrow \quad y_i[\omega * x+b]\geq 1-\boxed{\xi_i,} \quad i=1,\ldots,n$$

目标函数

$$\min \Phi(\omega)=\frac{1}{2}\|\omega\|^2$$

$$\Rightarrow \quad \frac{1}{2}\|\omega\|^2+\boxed{c\sum_{i=1}^{n}\xi_i}$$

c为惩罚函数，c越大表示对错误分类的惩罚越大

通过点积（内积）函数将非线性函数转换为线性函数

$$\boxed{K(x_i\cdot x_j)=\phi(x_i)\cdot\phi(x_j)}$$

拉格朗日乘子约束

$$Q(a)=\sum_{i=1}^{n}a_i-\frac{1}{2}\sum_{i,j=1}^{n}a_i a_j y_i y_j K(x_i\cdot x_j) \qquad \begin{cases} \boxed{0\leq a_i\leq C} \\ \sum_{i=1}^{N}a_i y_i=0 \end{cases}$$

线性分类函数

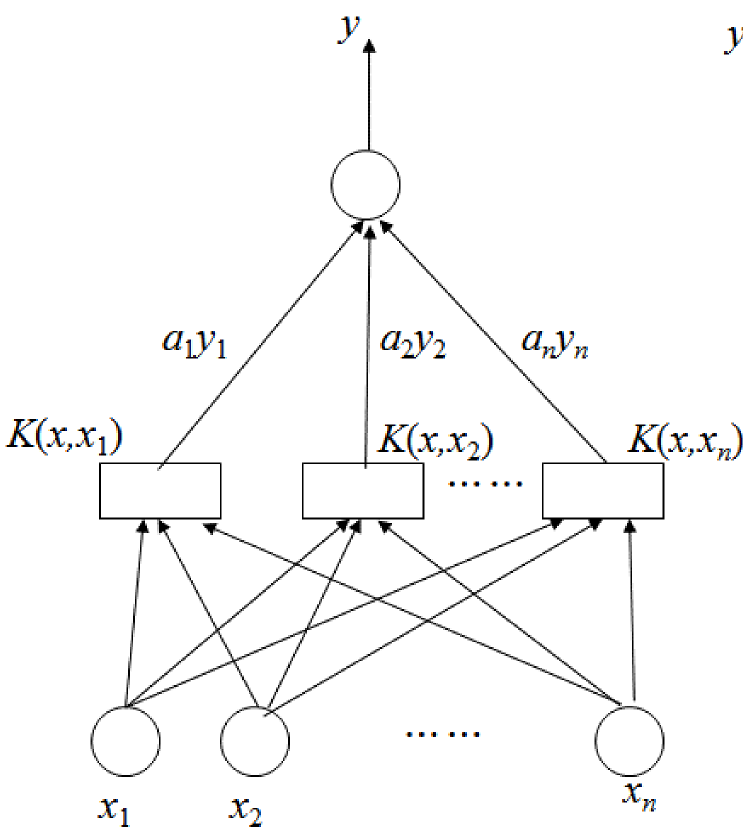
$$f(x)=\operatorname{sgn}\left\{\sum_{sv}a_i^* y_i(x_i,x)+b^*\right\}$$

非线性分类函数

$$\Rightarrow \quad f(x)=\operatorname{sgn}\{\sum_{i=1}^{n}a_i^* y_i K(x\cdot x_i)+b^*\}$$

广义最优分类面

- 非线性最优分类面
- 在形式上SVM分类函数类似于一个神经网络，输出是中间节点的线性组合，每个中间节点对应于一个支持向量
- 在SVM中，构造的复杂程度取决于支持向量的数目，而不是特征空间的维数



$y=\text{sgn}(\sum_{i=1}^{n} a_i\, y_i K(x, x_i) + b)$

权值：$w_i=a_iy_i$

基于n个支持向量的非线性变换（核函数）

输入向量：$x=(x_1,x_2,\ldots,x_n)$

核函数

- SVM的特点之一在于核函数的引入。低维空间向量集通常难于划分，解决的方法是将它们映射到高维空间，但会导致计算复杂度的增大，核函数可以巧妙解决这个问题。
- 采用不同的核函数构成不同的SVM，这样也形成了不同的算法，主要有以下几种形式：
- **(1) 线性函数** $\boxed{K(x, x_i) = \langle x, x_i \rangle}$
- **(2) 多项式函数** $\boxed{K(x, x_i) = \left[\langle x, x_i \rangle + 1\right]^d}$
- **(3) 高斯径向基核函数（Radial Basis Function, RBF）**

$$\boxed{K(x, x_i) = \exp\left\{-\frac{|x - x_i|^2}{\sigma^2}\right\}}$$

函数形式不确定情况下的选择

- **(4) 神经网络核函数（Sigmoid）** $\boxed{K(x, x_i) = \tanh\left[v\langle x, x_i \rangle + a\right]}$

SVM优点总结

- 具有良好的泛化能力，即**由有限的训练样本得到的小的误差能够保证使独立的测试集仍保持小的误差**
- 求解问题对应的是一个**凸优化问题**，因此局部最优解一定是全局最优解
- 核函数的成功应用，**将非线性问题转化为线性问题求解**
- 分类间隔的最大化，使得该算法具有较好的鲁棒性

小结

- 本讲小结
  - 理解分类与聚类、分类与数值预测的区别
  - 了解分类方法的评价指标
  - 掌握决策树归纳和支持向量机两种分类方法

- 推荐教材的相应章节
  - 《数据挖掘概念与技术》第八章（重点：8.1-8.2）

作业：关联

- 关联（根据教材177页6.6和6.7题改编）
  - 数据库中有5个事务。设min_sup=60%, min_conf=80%.

TID	购买的商品
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y}
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I, E}

  - 使用一种你熟悉的程序设计语言，如Python等，实现Apriori算法，并找出表中的频繁项集。
    - **要求：**不许用除默认库以外的包
  - 列举所有与下面元规则匹配的强关联规则（给出支持度s和置信度c）。其中X是代表顾客的变量，item$_i$是表示项的变量（如“A” “B”等）：
    - $\forall x \in transaction, buys(X, item_1) \wedge buys(X, item_2) \Rightarrow buys(X, item_3)\ [s, c]$
    - 如：$\{A, B\} \Rightarrow \{C\}\ [0.6, 1]$

作业：分类

- 分类（共享单车数据集）
- 将共享单车骑行数分为2类（与逻辑回归作业一致）
  - 不骑行：0；有骑行：1
- 构建决策树算法与SVM分类法，通过构建共享单车骑行分类模型
  - 不一定选择所有变量
  - 注意划分训练集与测试集（8:2）
  - 可以使用额外的包（如scikit-learn）
- 详细阐释每种方法中参数的设置及其理由
- 结合逻辑回归作业，对DT与SVM分类的结果进行分析
- 比较三个模型的优劣（如准确度、运算效率、鲁棒性等）

作业要求

- 关联：Apriori算法源代码
  - 发邮箱：945441387@qq.com（吕叶婷)
  - 截止时间：2022年5月6日09:59
    - 以邮件时间戳为准
  - 代码命名规则：学号_姓拼音_名拼音.扩展名
    - 例如：1953637_Sun_Rongjie.py
- 分类：纸质实验报告
  - 截止时间： 2021年5月27日上课前
  - 不要封面页，首页注明标题、姓名、学号
  - 宋体（英文、数字可以是Times New Roman）
  - 默认页边距、小四（12号）、1.5倍行距
  - 双面打印（节约纸），**不许超过10页**

第九讲 结束